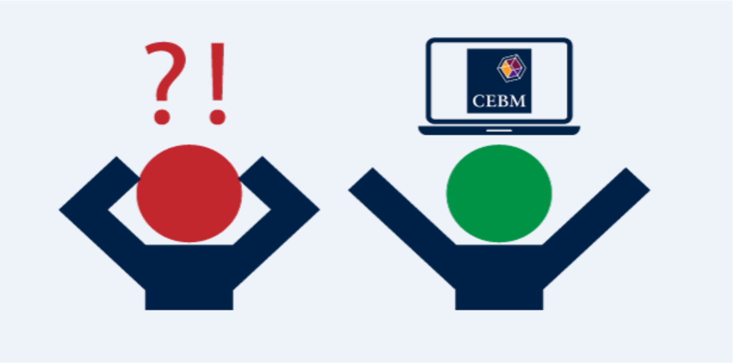
Tip for data extraction for meta-analysis – G2



How can you reduce the risk of errors and bias when extracting data?
Kathy Taylor

You may unintentionally make errors or introduce bias during data extraction. This will lower the quality of the meta-analysis and may lead to unreliable conclusions. There are a number of ways in which you can reduce the risk of data extraction errors and the risk of introducing potential biases:

1. **Make your data extraction efficient (post G1).** By doing so, you'll be less likely to get in a muddle and make errors.

2. **Extract the data twice, in pairs and independently,** compare the two sets of data and resolve any discrepancies, with a third person adjudicating if necessary. This is a well-established recommendation. If the outcomes are subjective, it may be necessary to have more than two people extract the data.

3. **Keep a copy of the data, as reported, untouched**. Perform your data extraction methods on a copy of the original data. If there are any queries later you can always go back to the original data without having to extract it again.

4. **If you have data queries, try contacting the authors.** You may want clarification about published data or enquire about data that you want but are not reported. When contacting authors it's important to emphasise that you're only asking for summary data that are readily available. Otherwise, authors may assume that you are inviting them to collaborate, which can lead to awkwardness. First email the corresponding author, and if they don't reply, try contacting the final author listed on the publication, as they're usually the principal investigator of the study.

5. **Check your units**. For example, in a review I worked on, some studies reported albumin excretion rate in mg/24hr and others reported it in µg/min. Data need to be converted to a common unit. As 1 mg=1000µg and 24 hours=24x60=1440 minutes so,
to convert mg/24hr to µg/min

   Multiply by $\frac{1000}{1440}$ because 1 mg/24hr = 1000 µg/24hr = $\frac{1000}{1440}$ µg/min

   and to convert µg/min to mg/24hr

   Multiply by $\frac{1440}{1000}$ because 1 µg/min = $\frac{1}{1000}$ mg/min = $\frac{1440}{1000}$ mg/24hrs

   Online convertors are very useful for finding out the conversion factors but, just like with calculators, you need to document your calculations. Also, note that conversion factors may vary according to what you're converting. For example,
1 mg/dL converts to 88.4 µmol/L for creatinine but 17.1 µmol/ L for bilirubin.
http://www.endmemo.com/medical/unitconvert

6. **Look out for dropouts**. When dealing with clinical trial data, for each treatment arm, you need to establish how many patients relate to the summary data that you've extracted. If you've extracted baseline data, then you can find the number of patients in each treatment arm in the CONSORT flow diagram. If you are extracting data for change from baseline or endpoint data you need to check for reports of the number of patients in each group who have not completed follow-up.

7. **Automate as much as you can**. Calculations coded (written) in a computer program can easily checked for errors and these calculations can be repeatedly rerun, without introducing human errors.

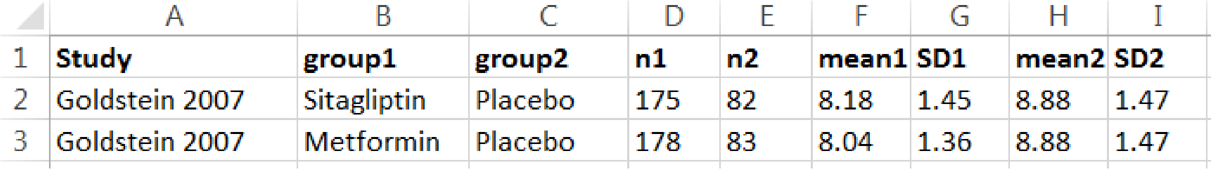
8. **Don't exclude studies just because the data you want are not reported**. Try to make sensible estimates from the given data so that as many studies as possible that meet your eligibility criteria are included in the meta-analysis. Follow my blog to find out how to convert data from what you're given into what you want.

9. **Be careful when there's more than one intervention group.** Sometimes a trial may report more than one intervention group and a control group. For example, in a trial of patients with type 2 diabetes Goldstein et al (2007) reported the effects of two glycaemic controlling drugs, Sitagliptin and Metformin, on HbAiC levels (Table). HbAiC levels are routinely measured in patients with diabetes to establish how well their diabetes is controlled.

Table. Two intervention groups

HbA1c	Placebo (n=165)	Sitagliptin (n=175)	Metformin (n=178)
Baseline	8.68 (1.00)	8.87 (0.99)	8.90 (1.00)
Week 24	8.88 (1.47)	8.18 (1.45)	8.04 (1.36)
Change from baseline	0.17 (0.00 to 0.33)	-0.66 (-0.83 to -0.50)	-0.82 (-0.98 to -0.66)

There are two sets of data for this trial – Sitagliptin vs Placebo comparison and Metformin vs Placebo comparison. You have to be careful to avoid double counting patients in the Placebo group. You need to split the patients in the Placebo group into two even (or approximately even) groups. One half then becomes the comparator group for the Sitagliptin vs Placebo comparison and the other half becomes the comparator group in the Metformin vs Placebo comparison. We assume that same summary data applies to each Placebo group. For example, allocating the extra patient (as 165 is an odd number) to the Metformin vs Placebo comparison would produce the following two sets of HbA1c data for week 24:



	A	B	C	D	E	F	G	H	I
1	**Study**	**group1**	**group2**	**n1**	**n2**	**mean1**	**SD1**	**mean2**	**SD2**
2	Goldstein 2007	Sitagliptin	Placebo	175	82	8.18	1.45	8.88	1.47
3	Goldstein 2007	Metformin	Placebo	178	83	8.04	1.36	8.88	1.47

10. **Look out for multiple reports from the same study.** These can be referred to as duplicate publications. They can range from reproductions of a published article, based on an identical population and outcomes (identical manuscript), to reports of subgroups or expanded populations and different outcomes. Including duplicate data can introduce considerable biases to your analysis. Often duplicate publications are covert, in that they don't cross reference the original study publication.

Overlapping patient populations may also occur in the case where data were originally collected for one study and then reanalysed in another study for a different purpose. For example, in a review of the prognostic value of 24 hour blood pressure variability, variability was measured in different ways (including standard deviation, night day ratio and morning blood pressure surge), and studies shared patient data to test the prognostic value of new measures. This review reported 36 different measures and five of the 24 included studies involved multiple publications of the same or overlapping patient populations. Data from duplicate publications were only included if the reported data were unique (e.g. about a different measure or outcome), and if two publications reported overlapping populations, the same measure and the same outcome, then the data for the larger population was included in the analysis. Ten publications arose from the International Database on

Ambulatory Blood Pressure Monitoring in Relation to Cardiovascular Outcomes (IDACO) group.

The recommendations for spotting overlapping data (duplicate publications) include comparing the author names, study locations and settings, population sizes, dates and study durations, and information about the study interventions. In the review mentioned above, it was also necessary to compare the blood pressure variability measures that were reported in each publication.

Data extraction often involves an element of ‘detective work’ to find the data you want whilst checking for covert duplicate publications and patient dropouts.

Here’s a tip…
Following best practice, being careful, and thinking like a detective will help minimise errors and reduce the possibility of introducing bias

My next blog post is going to be a video where I’ll show what I do when I find that data are only presented in graphical form. I’ll introduce you to some useful online data extraction software and also refer to some of the recommendations that I’ve given in this series on the practice of data extraction.

Dr Kathy Taylor teaches data extraction in Meta-analysis. This is a short course that is also available as part of our MSc in Evidence-Based Health Care, MSc in EBHC Medical Statistics, and MSc in EBHC Systematic Reviews.

Follow updates on this blog, related news, and to find out about other examples of statistics being made more broadly accessible on Twitter @dataextips